Annals of Clinical and Analytical Medicine

Original Research

Evaluation of the appropriateness and readability of chatgpt-4, claude, copilot and gemini responses to patient queries on eye health

Assessment of large language models in eye health

Meltem Toklu[1], Gizem Yel[2]

[1] Department of Ophthalmology, Ezine State Hospital, Çanakkale

[2] Department of Ophthalmology, Gazipasa State Hospital, Antalya, Türkiye

Abstract

Aim: This study aims to assess the appropriateness and readability of ChatGPT-4, Claude, Copilot, and Gemini's responses to frequently asked questions about eye health.

Material and Methods: Four commonly used Large Language Model (LLM) chatbots were asked 30 questions frequently asked by patients in an ophthalmology clinic. The appropriateness of the answers was evaluated by 2 experienced ophthalmologists. Readability was evaluated with 6 different indexes.

Results: The Likert scale (5-1) was utilized to determine adequacy. The responses generated by ChatGPT-4 had the highest average score of 4,90, followed by Anthropic Claude with 4,23, Google Gemini with 3,87, and last was Microsoft Copilot with a score of 3,50. None of the chatbot responses were rated as "very poor" or classified as incorrect. With an average Flesch-Kincaid Grade Level readability score of 7.60, only Anthropic Claude met the level of readability that The American Medical Association recommends. An assessment of the remaining five parameters revealed that the responses generated by all four chatbots exhibited readability levels exceeding that of a sixth-grade.

Discussion: In this study, ChatGPT-4 was more precise in responses, while Anthropic Claude scored better in terms of readability. These findings suggest that while LLMs may offer more comprehensive information, their complexity can hinder the average patient's comprehension.

Keywords

Ophthalmology, ChatGPT-4, Gemini, Copilot, Claude

DOI: 10.4328/ACAM.22622 Received: 2025-03-03 Accepted: 2025-04-07 Published Online: 06-05-2025 Printed: 2025-06-01 Ann Clin Anal Med 2025;16(6):445-449

Corresponding Author: Meltem Toklu, Department of Ophthalmology, Ezine State Hospital, Çanakkale, Turkiye.

E-mail: drmtunca@gmail.com P: +90 507 756 22 95

Corresponding Author ORCID ID: https://orcid.org/0000-0003-0890-324X

Other Authors ORCID ID: Gizem Yel, https://orcid.org/0000-0003-0071-9296

Introduction

The internet is an essential repository of health-related information for individuals pursuing knowledge about their medical conditions. Recent estimates suggest that approximately 80% of Americans utilize the Internet for health searches [1]. Searching for health information online is believed to positively impact health information consumers, as they are more likely to adhere to treatment after acquiring sufficient knowledge about their health conditions [2].

Over the last few years with the growing utilization of artificial intelligence (AI) natural language processing models have quickly gained popularity. Large language models (LLM) are AI systems capable of understanding and generating human language by processing vast amounts of text data. These systems can engage in dialogue-like interactions and operate in a manner akin to conventional existing search engines. Among various models, ChatGPT, or Chat Generative Pre-Trained Transformer, stands out as the most well-known. A variety of applications that have recently adopted LLMs include Google's Gemini, Microsoft's Copilot, and Anthropic's Claude. The training of these datasets utilizes a vast compilation of textual material, encompassing more than 400 billion words sourced from various online platforms, such as articles, books, and websites. The contribution of these LLMs has been remarkable for everyday users and the scientific research community. Numerous studies evaluated the effectiveness of chatbots in conducting medical assessments. For a significant number of individuals and their caregivers, these artificial intelligence tools function as a "remote health advisor." Advancements in AI chatbot technology, notably in LLMs, create promising avenues for improving patient education through online platforms. Therefore, these LLM chatbots are required to present reliable, understandable, and extensive replies. The variation among Google search outcomes, AI-driven chatbots, and the knowledge of medical professionals prompts a fresh dialogue regarding the caliber of information that patients might find on the internet.

Previous research indicates that ChatGPT's effectiveness in addressing patient inquiries varies within ophthalmology and medicine. Certain studies have revealed that ChatGPT's responses contained accurate information, whereas other research pointed out the existence of misleading and possibly harmful content in its replies [3-6]. Studies evaluating LLMs in different clinical conditions revealed that the readability of LLMs surpasses recommendations, requiring a higher level of patient understanding [7, 8].

While considerable proof emphasizes the effectiveness and possible implementations of ChatGPT in the medical domain, there is a significant gap in research focused on evaluating its competitors. Current research assessing the capabilities of large language models (LLMs) in the field of ophthalmology lacks studies that compare the quality of responses generated by ChatGPT with those from Claude, Gemini, and Copilot regarding general eye health. The objective of the present research is to evaluate the relevance and comprehensibility of the outputs produced by ChatGPT-4, Microsoft Copilot, Anthropic's Claude, and Google Gemini in response to comprehensive queries about eye health.

Material and Methods

This study focused on examining the relevance and readability of the content generated by LLM chatbots. A total of 30 questions emerged from discussions with clinicians, aimed at addressing the most prevalent concerns raised by patients in an ophthalmology setting. The phrasing of the questions employed terminology and language that a non-expert would understand, simulating the experience of patients interacting directly with the chatbots. Individual questions were directed at four frequently used AI chatbots: ChatGPT-4, Google Gemini, Microsoft Copilot, and Anthropic Claude. Researchers used each platform to start separate chat discussions to pose questions. Chatbot responses were recorded and extracted from their respective platforms and compiled into a list. These question-and-answer lists were then blinded and presented to the evaluators. The evaluators (MT, GY) graded the adequacy of AI-generated responses using a Likert scale (5-1): 5- excellent, 4-above average, 3-average, 2-below average, 1-very poor. Each grader had experience in clinical ophthalmology for ten years.

Readability was assessed using an online readability calculator with six different indices: Gunning Fog Index, Coleman- Liau Index, Flesch Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG) Index and Automated Readability Index. Every AI-generated response was evaluated and the means for each AI chatbot was calculated. The Gunning Fog Index assesses the occurrence of multi-syllable words in conjunction with the mean length of sentences. The simplicity and clarity of a given text are evaluated using this index score, which spans from 0 to 20. The Coleman Liau index uses sentences and letters as variables to evaluate the reading level of a text. It is used in addition to the other scores to assess particularly medical and law documents. FRES assigns a quantitative score ranging from 1 to 100, with a range of 60 to 70 generally regarded as satisfactory. For FRES a higher value corresponds with more readable text. The FKGL assigns a specific numerical score that reflects the corresponding educational grade level; for instance, a score of 8.2 suggests that a student reading at an 8th-grade level would be capable of comprehending the material. The SMOG score uses the frequency of polysyllabic words as a tool to assess the readability of text. It has been shown very useful in the healthcare sector, to evaluate patient comprehension of healthcare-related terms. Automated Readability Index relies on a factor of characters per word and produces an approximate representation of the US grade level needed to comprehend the text. The Coleman Liau Index, Automated Readability Index, and SMOG yield a score between 1 and 20 that is inversely proportional to the ease of readability of the text. The lower the score, the easier it is to read and comprehend the body of text [9]. Likewise, The Gunning Fog Index and FKGL reveal that a decrease in their respective values correlates with an increase in text clarity and readability.

Since the study was essentially a NonHuman Subjects Research, it was exempt from review by the Ethics Committee.

Statistical Analysis

Data were evaluated using the IBM SPSS Statistics Standard Concurrent User V 30 (IBM Corp., Armonk, New York, ABD) statistical package program. Descriptive statistics were given as several units (n), percentage (%), mean ± standard deviation, median, and interquartile distance values. The conformity of the data to normal distribution was evaluated with the Shapiro-Wilk normality test. The variance homogeneity of the chatbot groups was analyzed with the Levene test. For normally distributed data, Chatbot groups were compared with a one-way analysis of variance in repeated measures. For data that did not show normal distribution, Chatbot groups were compared using Friedman analysis. Bonferroni correction was made for all pairwise comparisons. A statistical significance threshold of p<0,05 was applied.

Results

Readability Assessments

Flesch-Kincaid Grade Level (FKGL) assesses grade level with a preferred grade reading level of 8. In these terms, according to Table 1, Anthropic Claude had the most preferred score with an average of 7,60, followed by Google Gemini with 9,05, ChatGPT-4 with 9,78, and Microsoft Copilot with 10.14. FKGL values were statistically different in chatbot groups (p<0.001). Anthropic Claude FKGL was statistically lower than other AI search engines. Google Gemini had FKGL statistically lower than Microsoft Copilot and ChatGPT-4. Analysis revealed that the scores for Microsoft Copilot and ChatGPT-4 were statistically comparable (Figure 1).

The Gunning Fog Index, SMOG Index, Automated Readability Index and Coleman-Liau Index scores of Anthropic Claude were statistically lower than other chatbots (p<0.001) (Figure 2). Flesch Reading Ease score for Anthropic Claude was statistically superior to other AI search engines. The Automated Readability Index score for Google Gemini was significantly lower than those of both Microsoft Copilot and ChatGPT-4, while ChatGPT's score also fell below that of Microsoft Copilot. Microsoft Copilot, Google Gemini, and ChatGPT-4 did not demonstrate any statistically significant difference in The Gunning Fog Index, Flesch Reading Ease, and Coleman-Liau Index scores. The SMOG Index score for Google Gemini demonstrated a

Table 1. Chatbot comparison

	Chatbot				Test statistics	
	Claude	Copilot	Gemini	GPT	Test value	p-value
Flesch-Kincaid Grade Level	7,60±1,38a	10,14±2,14b	9,05±1,23c	9,78±1,23b	30,046	<0,001†
Gunning Fog Index	9,09±1,71a	11,78±2,46b	11,02±1,33b	11,47±1,35b	21,591	<0,001†
Coleman-Liau Index	11,21 (1,72)a	13,17 (3,20)b	12,26 (2,61)b	12,33 (1,41)b	20,2	<0,001&
SMOG Index	9,80±1,19a	12,30±1,92b	11,46±1,14c	12,04±1,00b	34,042	<0,001†
Flesch Reading Ease	52,77 (10,04)a	51,04 (12,93)b	49,53 (10,53)b	49,04 (8,81)b	11,36	0,010&
Automated Readability Index	7,31 (2,07)a	10,85 (3,31)b	8,89 (2,18)c	9,65 (1,56)d	49,72	<0,001&

Data are given as mean±standard deviation or median (interquartile range) values. †: One-way repeated measures analysis of variance, &: Kruskal-Wallis test, a, b, c, and d superscripts indicate differences between chatbots at each row. There are no statistical differences between chatbots with the same superscripts.

Table 2. Chatbot comparison

	Chatbot				Test statistics	
	Claude	Copilot	Gemini	GPT	Test value	p value
Character Count	925,5 (191)a	881,0 (217)a	1079,5 (300)b	3227,5 (1303)c	61,84	<0,001&
Word Count	178,0 (29,0)a	169,0 (47,0)a	212,5 (51,0)b	666,0 (279,0)c	59,08	<0,001&
Likert scale	4,23±0,63a	3,50±0,73b	3,87±0,68c	4,90±0,31d	36,992	<0,001†

Data are given as mean±standard deviation or median (interquartile range) values. †: One-way repeated measures analysis of variance, &: Kruskal-Wallis test, a, b, c, and d superscripts indicate differences between chatbots at each row. There are no statistical differences between chatbots with the same superscripts.

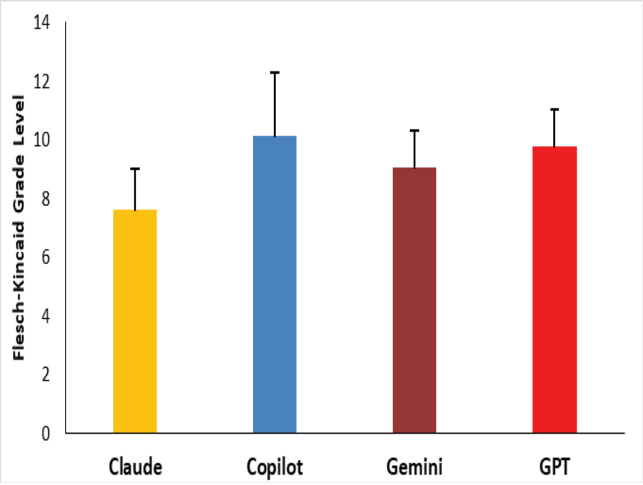

Figure 1. The level of large language model chatbots in terms of Flesch-Kincaid Grade Level is shown

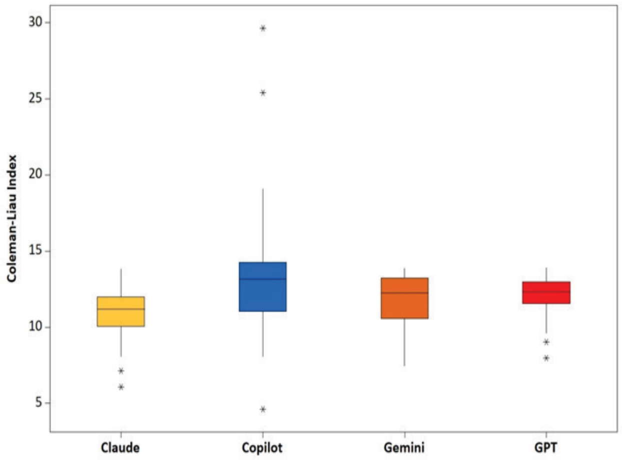

Figure 2. The scores of large language model chatbots in terms of Coleman- Liau Index shown on a boxplot

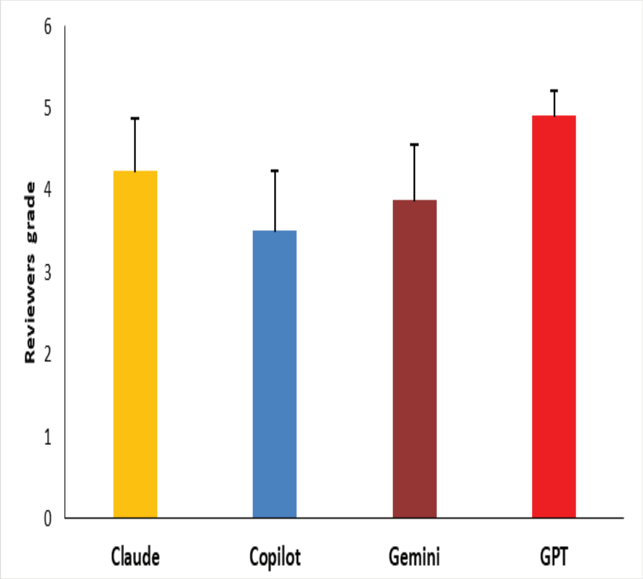

Figure 3. Likert scale mean score of chatbot responses to questions about eye health, as assessed through ratings by two expert ophthalmologists

statistically significant reduction compared to both Microsoft Copilot and ChatGPT-4, whereas the scores for Microsoft Copilot and ChatGPT-4 were not statistically significant.

Analysis of the counts of words and characters revealed that ChatGPT-4 exhibited markedly greater figures compared to the other LLM chatbots, with a significance level of p metrics (Table 2). The analysis of word and character count metrics for Google Gemini demonstrated markedly elevated values compared to those observed for Anthropic Claude and Microsoft Copilot, which exhibited no statistically significant differences between them.

Adequacy of Information

The Likert scale was utilized to determine adequacy. The responses generated by ChatGPT-4 had the highest average score of 4,90, followed by Anthropic Claude with 4,23, Google Gemini with 3,87, and last was Microsoft Copilot with a score of 3,50 (Table 2). Google Gemini had a statistically higher score than Microsoft Copilot. Anthropic Claude had a score statistically higher than Google Gemini and Microsoft Copilot. The performance of ChatGPT-4 surpassed that of its counterparts, demonstrating a statistically significant score. (Figure 3).

Discussion

This research aimed to evaluate how various language models respond to frequently asked questions about prevalent eye disorders. Our research revealed that responses generated by ChatGPT-4 were rated as more accurate and were preferred by clinicians. In terms of adequacy, ChatGPT-4 had the highest average score of 4,90, followed by Claude with 4,23. The evaluation of the responses indicated that none were rated as "very poor" or classified as incorrect. It was noted that ChatGPT generated appropriate replies that comprehensively responded to all queries, using bullet points to facilitate clarity. Claude offered responses that met expectations, regularly using a bullet-point format to express ideas clearly. In all responses, Claude recommended that individuals seek the expertise of qualified medical professionals to properly assess their medical conditions. Among large language models, CoPilot uniquely features hyperlinks that serve as citations in its replies. A significant number of hyperlinks led users to non-academic sites, which compromised the reliability of the information presented.

The American Medical Association (AMA) advises that online patient education materials should be written at or below a 6th-grade reading level to ensure understandability among patients with lower health literacy [10]. Readability refers to how easily a reader can comprehend a written text, in this study six different indices were used for evaluation. The FKGL measure functions as an alternative indicator for determining the readability associated with specific grade levels. With an average FKGL readability score of 7.60, Anthropic Claude meets the level of readability that the AMA recommends. An assessment of the remaining five parameters revealed that the responses generated by all four chatbots exhibited readability levels exceeding that of a sixth-grade education potentially limiting their accessibility to a broader patient demographic.

While a recent study determined that ChatGPT produces incomplete and inaccurate information about common ophthalmic conditions our results support two prior studies which found that ChatGPT was largely effective in responding to patients' eye care questions [5, 11, 12]. Research that compared Google and ChatGPT indicates that responses generated by ChatGPT were more accurate but written at a significantly higher grade level than responses generated by Google [12, 13]. In a different analysis examining ChatGPT's communication skills with individuals diagnosed with glaucoma, researchers found that the model proficiently offers general information, encompassing definitions, and potential therapeutic strategies. However, the readability scores were high [14]. Ichhpujani et al compared the appropriateness and readability of Google Bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma, they found that the answers generated by ChatGPT-3.5 are more accurate than the ones given by Google Bard but reported that comprehension of ChatGPT-3.5 answers may be difficult for the public with glaucoma [15]. According to Guven et al., the responses produced by ChatGPT-4 were considered satisfactory, yet the complexity of these responses made them challenging to read in the context of frequently posed questions about strabismus and amblyopia [16]. In a recent study appropriateness and readability of responses provided by ChatGPT-3.5, Bard, and Bing Chat to frequently asked questions about keratorefractive surgery were assessed, it was reported that ChatGPT-3.5 had the highest accuracy but the readability scores were more challenging than the recommended level [17].

Limitation

The main limitation of our study was that it involved a small number of questions. Furthermore, we employed subjective grading to evaluate the appropriateness of the responses, which means our findings might not apply to other chatbots and similar technologies.

Conclusion

In this study, ChatGPT-4 was more precise in responses, while Anthropic Claude scored better in terms of readability.

These findings suggest that while LLMs may offer more comprehensive information, they do not always improve accessibility for the average patient. LLMs represent an accurate information source for patients and can be utilized by providers as a patient educational tool but presently it does not meet the necessary standards without concurrent supervision from healthcare providers. Furthermore, acknowledging that the success of knowledge relies on the comprehension skills of the patient is crucial, further studies using different prompts and evaluation methods will be needed to better assess the accuracy and understandability of the information provided by LLMs in ophthalmology and other fields of medicine.

Scientific Responsibility Statement
The authors declare that they are responsible for the article's scientific content including study design, data collection, analysis and interpretation, writing, some of the main line, or all of the preparation and scientific review of the contents and approval of the final version of the article.

Animal and Human Rights Statement
All procedures performed in this study were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or compareable ethical standards.

Funding: *None*

Conflict of Interest
The authors declare that there is no conflict of interest.

References

1. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online Health Information Seeking Among US Adults: Measuring Progress Toward a Healthy People 2020 Objective. Public Health Rep. 2019;134(6):617-25.
2. Jia X, Pang Y, Liu LS. Online Health Information Seeking Behavior: A Systematic Review. Healthcare (Basel). 2021;9(12):1740.
3. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. Aesthetic Plast Surg. 2023;47(5):1985-93.
4. Samaan JS, Yeo YH, Rajeev N, Hawley L, Abel S, Ng WH et al. Assessing the Accuracy of Responses by the Language Model ChatGPT to Questions Regarding Bariatric Surgery. Obes Surg. 2023;33(6):1790-6.
5. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. Eye (Lond). 2024;38(7):1368-73.
6. Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: A scoping review of applications in medicine. Front Med. 2024;11: 1477898.
7. Mu X, Lim B, Seth I, Xie Y, Cevik J, Sofiadellis F, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI, and ChatGPT. Skin Health Dis. 2023;4(1):e313.
8. Seth I, Lim B, Xie Y, Cevik J, Rozen WM, Ross RJ, et al. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. Aesthet Surg J Open Forum. 2023;5:ojad084.
9. Robinson E, McMenemy D. To be understood as to understand: A readability analysis of public library acceptable use policies. J Librariansh Inf Sci. 2020; 52(3):713-25.
10. Weiss, Barry D. Health literacy. Am Med Assoc. 2003; 253(3):358.
11. Bernstein IA, Zhang YV, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. JAMA Netw Open. 2023;6(8):e2330320.
12. Cohen SA, Brant A, Fisher AC, Pershing S, Do D, Pan C. Dr. Google vs. Dr. ChatGPT: Exploring the Use of Artificial Intelligence in Ophthalmology by Comparing the Accuracy, Safety, and Readability of Responses to Frequently Asked Patient Questions Regarding Cataracts and Cataract Surgery. Semin Ophthalmol. 2024;39(6):472-9.
13. Cohen SA, Yadlapalli N, Tijerina JD, Alabiad CR, Chang JR, Kinde B, et al. Comparing the Ability of Google and ChatGPT to Accurately Respond to Oculoplastics-Related Patient Questions and Generate Customized Oculoplastics Patient Education Materials. Clin Ophthalmol. 2024;18: 2647-55.
14. Wu G, Lee DA, Zhao W, Wong A, Sidhu S. ChatGPT: Is it good for our glaucoma patients? Front Ophthalmol. 2023;3: 1260415.
15. Ichhpujani P, Parmar UPS, Kumar S. Appropriateness and readability of Google Bard and ChatGPT-3.5 generated responses for surgical treatment of glaucoma. Rom J Ophthalmol. 2024;68(3):243-8.
16. Guven S, Ayyildiz B. Acceptability and readability of ChatGPT-4 based responses for frequently asked questions about strabismus and amblyopia. J Fr Ophtalmol. 2024;48(3):104400.
17. Doğan L, Özer Özcan Z, Edhem Yılmaz I. The promising role of chatbots in keratorefractive surgery patient education. J Fr Ophtalmol. 2025;48(2):104381.

How to cite this article:
Meltem Toklu, Gizem Yel. *Evaluation of the appropriateness and readability of chatgpt-4, claude, copilot and gemini responses to patient queries on eye health. Ann Clin Anal Med 2025;16(6):445-449*